A hierarchical regression mixture model for inferring gene regulatory networks

Mayetri Gupta
gupta@bios.unc.edu

University of North Carolina at Chapel Hill
Fundamental question: how can we understand the biological mechanisms leading to disease?

Co-regulated genes share similar upstream patterns

Identify genes that are differentially expressed under different treatments or conditions
Proteins bind to DNA to activate gene transcription

Position specific weight matrix (PSWM) or Motif
Gene regulation in complex genomes

- **Harder problem:** many transcription factors working in co-ordination
- **LARGE sequence search space:** using sequence data only → many false positives?
Upstream regulation ↔ Downstream expression

- Gene expression contains information about sequence motifs
- Sequence may contain information on gene co-regulation

Expression clustering ➔ Motif discovery

or

Expression clustering ← Motif discovery ?

What if initial clustering is inaccurate?
Measurements over 18 time points, 3 different experiments
\(~800\) genes known to be cell-cycle dependent

Do clusters of genes share common TFs?
Do certain TFs work combinatorially on groups of genes?

*color: time when gene is active*
REDUCER (Bussemaker, Nat. Genet. 2001) correlates expression of gene with number of motif occurrences

MDScan (Liu, Nat Biotech. 2002) Most strongly differentially expressed genes → candidate motifs.

Motif Regressor (Conlon, PNAS 2003)

Multiple regression model: Sum of motif effects explains gene expression

\[ Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + \varepsilon_g \]

- \( Y_g \): expression of gene \( g \);
- \( S_{mg} \): motif-match score
Non-parametric approaches

- Phuong et al. (Bioinformatics, 2004): Classification Trees (CART)
  
  *Arbitrary decision criterion based on the number of occurrences of a motif type*

- Multivariate adaptive regression splines (Das et al, PNAS 2004)

*Joint sequence-expression model without parametric connection*

- Holmes and Bruno (ISMB proc., 2000) joint likelihood for sequence and expression data
Using motif information in gene clustering

Infer sets of transcription factors involved in regulating groups of genes

- Higher transcriptional activity $\rightarrow$ greater presence of TF binding sites, more pronounced expression changes
- Genes within a “cluster” may be correlated, with or without sharing common transcription factors
- Measurements on the same gene in different conditions may be correlated due to sharing the same upstream transcription factor binding sites
Linear mixed effects model

\[ y = X\beta + Zb + \varepsilon \]

- \( \beta \): fixed effects
- \( b \): random effects
- \( \varepsilon \sim N(0, \tau_0^{-1}I) \)

- \( y \): gene expression
- **Fixed effects: Sequence Motif**
  Levels of factor are reproduced exactly if experiment is repeated
- **Random effects: Expression cluster**
  Levels of factor (expression + clusters) may not be reproduced exactly if experiment is repeated
Joint Model for Sequence-Expression

Complication: gene cluster identity cannot be assumed known
Z matrix not “fixed”

Conditional on cluster $k$, $(k = 1, \ldots, K)$, vector of log-expression
values of gene $g$ generated from mixed-effects model:

$$Y_g|z_g = k, X, \text{ parameters} \sim N(\xi_g + X_g^T \beta_k 1, \sigma^2_k I) \quad (\equiv f_k)$$
Joint Model for Sequence-Expression

Complication: gene cluster identity cannot be assumed known
$Z$ matrix not “fixed”

Conditional on cluster $k$, $(k = 1, \ldots, K)$, vector of log-expression values of gene $g$ generated from mixed-effects model:

$$ Y_g | z_g = k, X, \text{ parameters } \sim N(\xi_g + X_g^T \beta_k 1, \sigma_k^2 I) \quad (\equiv f_k) $$

Unconditionally, a mixture model

$$ P(Y | X, \text{ parameters}) = \prod_{\text{genes } g} \left[ \sum_{\text{cluster } k} \pi_k f_k(Y_g | \text{ parameters}) \right] $$

Which $\beta$'s significant, in which cluster?
Bayesian hierarchical formulation

Prior distributions for cluster $k$ parameters:

“Expression” model:

$$
\mu_k \sim N(\cdot, \nu_{k0}^2 I)
$$

$$
\sigma_k^2 \sim \text{InvGamma}(\cdot, \cdot)
$$

$$
\xi_g | z_g = k \sim N(\mu_k, \tau_0 \sigma_k^2 I)
$$

“Sequence” model:

$$
\beta_k \sim N(\beta_0, V_k)
$$

Probabilities of cluster membership:

$$
P(z_g = k) = \pi_k
$$

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)
$$

$\mathbf{G}$ genes; $\mathbf{T}$ measurements per gene
For each $\beta_k$, we use multivariate extension of g-prior (Zellner, 1986), so that

$$V_k = \frac{c\sigma_k^2}{T} S_k^{-1},$$

where $S_k = \sum_{z_g=k} X_g X_g^T$.

Why use g-prior?

- Computational efficiency, varying $c \rightarrow$ more/less informative
- Induces dependence among genes in a cluster due to sequence effects

$$\text{Cov}(Y_g, Y_{g'} | Z_g, Z_{g'} = k) = v_{k0}^2 I + \frac{c\sigma_k^2}{T} 1 [X_g^T S_k^{-1} X_g] 1^T$$
Every non-site position multinomial with
\[ \theta_0 = (\theta_{01}, \ldots, \theta_{04}) \]

Every motif position \( i \) multinomial with
\[ \theta_i = (\theta_{i1}, \ldots, \theta_{i4}) \]

Product Multinomial model

Challenge: Find position of sites and \( \theta \)'s


Sequence Motif Scoring

Starting set of motifs

- **De-novo**: Different clusters of genes exhibiting “strong” up- or down-regulation (MDScan)
- **Databases**: Derived from experimental data

Motif score for \( w \)-width motif \( j \) and upstream sequence \( g \):

\[
X_{gj} = \sum_{\text{positions } i} \frac{P(\text{seq}(i, i + w - 1) | \text{motif } j)}{P(\text{seq}(i, i + w - 1) | \text{null})}
\]
Sequence motif selection

Initial set of $D$ motif candidates ($D$ can be large!)
In regression model, want to know which motifs correlated “significantly” with response (gene expression)

$$u = (u_1, \ldots, u_D)$$ where

$$u_j = \begin{cases} 
1 & \text{if motif } j \text{ is in model} \\
0 & \text{otherwise.} 
\end{cases}$$

Prior probability of motif inclusion

$$P(u) = \prod_{j=1}^{D} \eta^{u_j} (1 - \eta)^{1-u_j}$$

Variable selection from LARGE potential set
Outline of method

select motifs from large initial set

update clusters given motifs active in each class

update model parameters given cluster membership and functional motifs in each class

Complications

- Cluster membership $z$ unknown
- Number of clusters $K$ may be unknown (assume fixed for now)
- Number of motif candidates is large
Parameter updating using MCMC

Update from joint posterior distribution

\[ P(\theta, \beta, u, z|Y, X, K) \]

\[ \theta = (\mu, \sigma^2, \pi) \]

- For updating steps for parameters \( \theta, \beta \) and \( z \), marginalize over other parameters for efficiency (Conjugate forms permit this)
- For updating \( u \), use evolutionary Monte Carlo (Liang and Wong, JASA 2001)

Select motifs that have most effect on expression, and differentiate most among clusters
Three simulated data sets

Regression coeffs. for motifs corresponding to 3 PSWMs from JASPAR database \textit{SAP}1, \textit{SRF}, and \textit{MEF}2

200 "genes" in \( K = 2 \) clusters

2 measurements each

\begin{array}{c|ccccc}
\text{Data 1} & \text{Data 2} & \text{Data 3} \\
\hline
\text{Coeff.} & C_1 & C_2 & C_1 & C_2 & C_2 \\
\beta_{M_1} & 2 & 0 & 2 & 0 & -2 \\
\beta_{M_2} & 0 & 2 & 0 & -2 & 0 \\
\beta_{M_3} & 0 & 0 & 0 & 0 & 0 \\
\end{array}

(Motif 3 not present in data)

\( Y_1 \) with motif 1, 2, 3 scores (columns)

in 3 data sets (rows)
Simulation study results: Bayes factors

Optimal choice: $K = 2$

Marginal model probability for $M_K$ through Double mixture importance sampling

$$P(Y|M_K) \overset{\triangle}{=} \frac{1}{N_t N_s} \sum_t \sum_s P(Y|Z^{(t)}, \theta^{(s)}, K) \frac{\pi_1(\theta^{(s)}|z^{(t)}) \pi_2(z^{(t)})}{f_1(\theta^{(s)}|z^{(t)}) f_2(z^{(t)})}$$
Simulation: $\beta$ estimates

Data 1

Data 2

Data 3

Data sets 1 and 2: Motifs 1,2 selected
Data set 3: Motif 1 selected
Pick different groups of genes that are highly differentially expressed

Sets of motif candidates of widths 7-12 bp (using MDscan)

Motif overlaps lead to collinearity: remove motifs with correlation > 0.5 with a higher-ranked one → 32 candidate motifs

Two consecutive time points: 1-2, 3-4, ..., 17-18
Number of clusters $K^*$

$\log(\text{BF})$ compared to 1-component model

<table>
<thead>
<tr>
<th>Interval</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K^*$</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Significant motifs over time intervals

- SCB
- MCB
- SFF
- MCM1
- MCB
- MCB
- MCB
- SFF
- MCB
- SFF

Significant motifs over time intervals

- SCB
- MCB
- SFF
- MCM1
- MCB
- MCB
- SFF

Significant motifs over time intervals
Motif influence at different time intervals

Motif index

Selected motif types for 9 time intervals for optimal $K$
## Significant motifs match experimental PSWMs

<table>
<thead>
<tr>
<th>Index</th>
<th>TF name</th>
<th>Consensus</th>
<th>Expt.</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>MCM1</td>
<td>CGAAGAG/CTCTTCG</td>
<td>CCNNNWWRRGG</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>GCN1</td>
<td>TCAGTCA/TGACTGA</td>
<td>TCAGTCA</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>[CSRE]</td>
<td>GGACAGA/TCTGTCC</td>
<td>[YCGGAYRRRAWGG]</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>MCB</td>
<td>ACGCGTA/TACGCGT</td>
<td>WCGCGW</td>
<td>G1</td>
</tr>
<tr>
<td>16</td>
<td>SFF</td>
<td>AACAACA/TGTTGTT</td>
<td>GTMAACAA</td>
<td>M</td>
</tr>
<tr>
<td>18</td>
<td>MCM1</td>
<td>CCAATTAGG/CCTAATTGG</td>
<td>CCNNNWWRRGG</td>
<td>M</td>
</tr>
<tr>
<td>20</td>
<td>[RME1]</td>
<td>TTCAGGTAC/GTACCTGAA</td>
<td>[GAACCTCAA]</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>SCB</td>
<td>CGCGAAAAA/TTTTTTCGCG</td>
<td>CNGCDDD</td>
<td>G1</td>
</tr>
<tr>
<td>25</td>
<td>PHO4</td>
<td>CGTACGTAC/GTACGTACG</td>
<td>CACGTA</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>GCAACATC/GATGCGAAG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
K \equiv \{G \text{ or } T\};
M \equiv \{A \text{ or } C\};
N \equiv \{A \text{ or } C \text{ or } G \text{ or } T\};
R \equiv \{A \text{ or } G\};
W \equiv \{A \text{ or } T\};
Y \equiv \{C \text{ or } T\}
\]
Treating gene expression clustering as a variable may help in discovering relationships between functional sequence motifs, and groups of genes they regulate.

- Different groups of genes may behave as a cluster at different time points.
- Upstream sequence motifs “constant” but effects/interactions over time may vary.
Further extensions

- Motif scoring issues: sensitivity, co-occurrence of sites
- Efficient model selection
- Extension to high density ChIP tiling arrays

Acknowledgement:
Joseph G. Ibrahim (UNC), Jason Lieb (UNC)

UNC high-performance scientific computing group
Model Selection: number of clusters $K$

Likelihood-based methods not valid (BIC, etc.)

**Bayes factor:** ratio of model marginal probabilities

Marginal probability for model $M_K$

$$P(Y|M_K) = \sum_z \int_{\theta} P(Y|Z,K)p(\theta|z)p(Z|K)d\theta$$

*Double mixture importance sampling*

$$P(Y|M_K) \overset{\sim}{=} \frac{1}{N_tN_s} \sum_t \sum_s P(Y|Z^{(t)},\theta^{(s)},K) \frac{\pi_1(\theta^{(s)}|Z^{(t)})}{f_1(\theta^{(s)}|Z^{(t)})} \frac{\pi_2(Z^{(t)})}{f_2(Z^{(t)})}$$
Model Selection: Number of clusters

Double mixture importance sampling

\[
P(Y|M_K) \triangleq \frac{1}{N_t N_s} \sum_t \sum_s P(Y|Z^{(t)}, \theta^{(s)}, K) \frac{\pi_1(\theta^{(s)}|z^{(t)}) \pi_2(z^{(t)})}{f_1(\theta^{(s)}|z^{(t)}) f_2(z^{(t)})}
\]

Challenge: Good sampling densities \( f(\cdot) \) for \( z \) and \( \theta \)

For a simpler case (\( \theta \) marginalized), Raftery et al (TR, 2003) propose permutation-based methods to find good candidate sampling densities for \( z \).